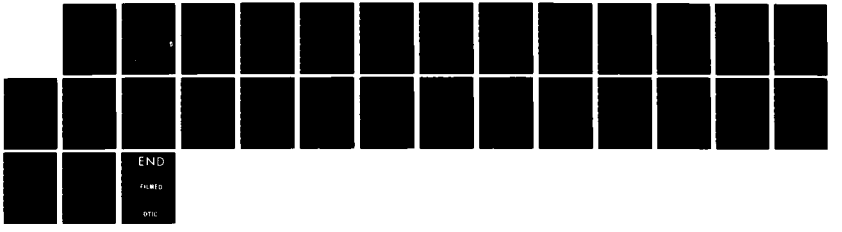
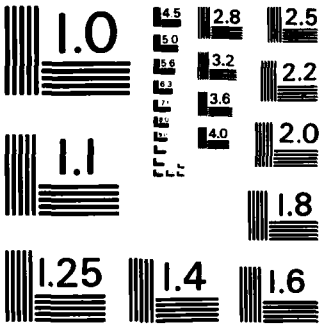
AD-A158 546 RECURSIVE PARTITIONING USING RANKS(U) STANFORD UNIV CA 1/1
LAB FOR COMPUTATIONAL STATISTICS M R SEGAL AUG 85
LCS-TR-15 N00014-83-K-0472

UNCLASSIFIED F/G 12/1 NL



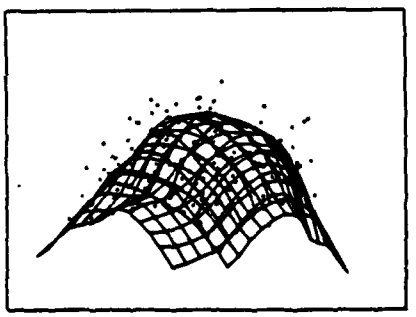


RECURSIVE PARTITIONING USING RANKS

Mark Robert Segal

Technical Report No. 15
August 1985

Laboratory for Computational Statistics



Department of Statistics
Stanford University

RECURSIVE PARTITIONING USING RANKS

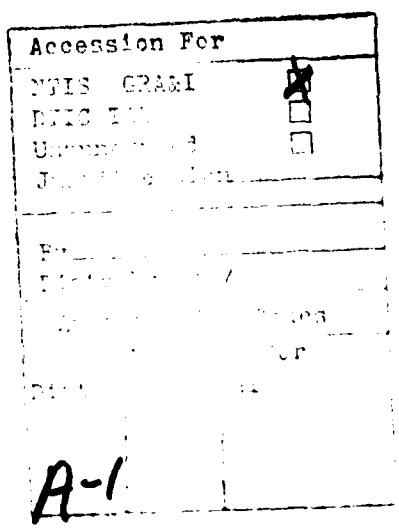
Mark Robert Segal

Department of Statistics

Stanford University

Abstract

Replacing the conventional splitting rules used in constructing regression trees by rules based on two sample rank statistics affords many advantages and equally poses some problems. Among the former are (1) computational ease, (2) invariance under monotone transformations of the response and (3) worthwhile extension to censored data. The difficulties involve devising good pruning strategies in the absence of within node loss. These are addressed using look-ahead, bottom-up techniques. Some real-world and simulation performances of the methodology are presented.



1. Introduction.

This paper proposes some modifications to the conventional regression-tree methodology with the primary motivation of facilitating an extension to censored data. However, the suggested changes have merit in their own right and comparisons with the existing techniques plus some additional extensions are also presented. The basic alteration is the replacement of the goodness-of-split criterion with a measure of node *separation* as opposed to within node homogeneity. The measures used are various two-sample statistics (principally rank statistics). Their introduction further necessitates changing the pruning algorithm used to determine desirable tree size. The next section is a brief overview of current regression-tree (or recursive partitioning) methodology. Section three deals with the new splitting criteria and in particular addresses the censored data issue. Section four indicates how the new pruning strategies work. The fifth section comprizes some examples, both real-world and simulated, and the sixth discusses properties and potential of the new approach. A means whereby tree-structured techniques can be used in multi-response situations is also proposed.

Repeated allusion is made to the definitive reference "Classification and Regression Trees" by Breiman, Friedman, Olshen and Stone (1984), which is referred to as CART, and in which stand alone section numbers should be sought.

2. Regression Tree Methodology.

A simplified description of regression-trees is presented in this section, so that the subsequent reformulations can be understood. Attention here is restricted to the familiar regression setting — there are p predictor variables $X_1, X_2, \ldots, X_p$ and a (continuous) response Y. No comment is made with respect to issues such as the treatment of missing values (§5.3.2), or variable importance (§5.3.4) for which carry over from the standard methods is straightforward. However, to the extent that neither ranks nor censoring pertain to categorical responses, no extensions to classification trees or class-probability trees (§4.6) exist.

In order to construct a regression-tree four constituent components are required. These are:

1. A set of (binary) questions of the form

"Is $\vec{x} \in A$?" where $A \subset X$, the predictor space.

The answer to such a question induces a partition, or split, of the predictor space. The subsample associated with region A is called a *node*.

2. A goodness-of-split criterion $\phi(s,t)$ that can be evaluated for any split s of any node t. The criterion is used to assess the worth of the competing splits, where (in CART) worth pertains to within node homogeneity.
3. A means for determining the appropriate tree size.
4. Summaries for the terminal nodes of the selected tree.

What follows is an elaboration of these aspects.

The number of possible splits in 1 above, is reduced to a computationally feasible number by constraining that:

(a) each split depends upon the value of only a single predictor variable [note: this restriction can be loosened; the software (CART^{TM}, 1984) permits splits on *linear* combinations of predictors].

(b) for ordered predictors X_j, only splits resulting from questions of the form "Is $X_j \leq c$?" are considered.

(c) for categorical predictors all possible splits into disjoint subsets of the categories are allowed.

The tree is grown as follows: for each node (the initial or *root* node comprises the entire sample)

1. examine every allowable split on each predictor variable.
2. select and execute (create two new children nodes) the *best* of these splits.

Steps 1 and 2 are then reapplied to each of the children nodes, and so on.

"Best" in 2 above, is assessed in terms of the goodness-of-split criterion. Two such criteria are espoused in CART and available in the associated software. These are *Least Squares* (LS) §8.3, 8.4 and *Least Absolute Deviations* (LAD) §8.11. Both afford a comparison based on subadditive "between/within" decomposition, where between alludes to the homogeneity or loss measure applied to the parent node. This paper is concerned with the replacement of these

by various two-sample statistics, that detect node separation and why such a substitution is warranted. For point of reference and specificity the definition of the LS criterion is presented here. The obvious changes give rise to LAD (or any other between/within criterion such as is used in §6.4).

Let t designate a node of the tree. That is, t contains a subsample $\{(\vec{x}_n, y_n)\}$. Let $N(t)$ be the total number of cases in t and let

$$\bar{y}(t) = \frac{1}{N(t)} \sum_{\vec{x}_n \in t} y_n$$

be the node response average. Then the within node sum-of-squares is given by

$$SS(t) = \sum_{\vec{x}_n \in t} (y_n - \bar{y}(t))^2$$

Now suppose a split s partitions t into left and right daughter nodes t_L and t_R. The LS criterion is

$$\phi(s,t) = SS(t) - SS(t_L) - SS(t_R)$$

and the best split s^* is the split such that

$$\phi(s^*,t) = max\ \phi(s,t)$$

where the maximum is taken over all permisible splits s.

A LS regression-tree is constructed by recursively splitting nodes so as to maximize the above ϕ function. The criterion is such that we create smaller and smaller nodes of progressively increased homogeneity. It remains unresolved as to what constitutes an appropriate sized tree Originally (the AID program, Morgan and Sonquist (1963)) this was determined by use of stopping rules: if a node became too small or the improvement ($\phi(s^*,t)$) resulting from the best split was not sufficient (to surmount some preset threshold), then the node was declared *terminal*. This proved unsatisfactory on account of the rigidity of the thresholds. In some instances overfitting via too large a tree would occur. In others, underfitting would result from rejection of a split precluding the emergence of subsequent worthwhile splits. There is an analogy to step-wise versus subset regression in terms of failure to capture important interactions. The problem was redressed by

1. initially growing a very large tree
2. iteratively *pruning* this tree all the way back up (to the root node), thereby creating a nested sequence of trees
3. selecting the best tree from this sequence using cross-validation

This procedure is detailed in Ch.3 CART. The means for performing the pruning in 2 is called "minimal cost-complexity pruning" §3.3. This paper presents some alternatives to this when no within node cost is available (for LS the within node cost is just $SS(\cdot)$).

The fourth item - summaries for the terminal nodes - is not dwelt on here. The point behind such measures is primarily predictive. In the LS and LAD situations the node mean and median respectively constitute natural measures. However, for splitting based on two-sample statistics, no such quantities arise. Thus, it is left as a user-specified option as to how a terminal node is summarized. More on this is said when censored data is discussed. Bloch and Segal (1985) deal with *classification*-trees in a setting where non-standard statistics are used.

3. Two Sample Statistic Splitting.

Instead of gearing our splitting criteria to optimizing within-node homogeneity, we could reward splits that resulted in large between-node separation. The magnitude of any two-sample statistic affords such a goodness-of-split measure. Such a change constitutes more than just a rephrasing of the problem. Whilst it is generally the case that splitting based on two-sample t statistics with unpooled variance estimates gives results strongly resembling those obtained from LS splitting as outlined in section 2, there is no algebraic equivalence and problems can be contrived where results are dissimilar.

The fact, that in all the cases analysed, splitting using t statistics and splitting using LS gave comparable results, supports the usage of two-sample statistics: given that the two techniques produce analogous results *and* LS gives worthwhile answers the new approach must be doing something reasonable. But why replace a proven method with one that is harder to motivate and offers no computational savings? The answer lies in the advantages provided by using two-sample *rank* statistics (TSRS). These include all the conventional desiderata of

ranks plus some additional benefits:

1. Invariance under monotone transformation of the response Y. The regression-trees created by using LS or LAD possessed such invariance only with respect to monotone transformations of the (ordered) predictors $X_1, X_2, \cdots, X_p$. This means, for instance, that the optimal split is the same regardless of whether we use X_1 or $\tilde{X}_1 = g(X_1)$ for some monotone g. If the optimal split on X_1 is $X_1 \leq c$ then the optimal split using $\tilde{X}_1$ will be $\tilde{X}_1 \leq g(c)$ (§2.7). However, it is only through the use of TSRS that such properties will hold under monotone transformations of Y to $\tilde{Y} = h(Y)$. This is clearly a worthwhile property when there is no natural response scale in which to work.

2. Insensitivity to outliers in the response space. The use of LS, and to a lesser extent LAD, is subject to the familiar sensitivity to extreme Y observations. This, in the regression-treesetting, is not necessarily a drawback, since such outliers will be isolated into their own (single case) terminal nodes. Still, the influence on overall tree topology can be distorting and the interpretation of splits leading to the isolation of the outlier can be problematic. Friedman (1979) regards the presence of outliers as weakening the LS procedure by wasting splits. Using TSRS amounts to a down-weighting of extrema. It would equally be possible to achieve this end by using resistant averages. Indeed, it is found that the trees resulting from LAD splitting agree far more closely with those obtained from using TSRS splitting than do the corresponding LS trees. However, it is computationally easier and perhaps more natural to employ ranks.

   Note: none of the methods address outliers or homogeneity or clustering in the *predictor* space.

3. Computational feasibility. The actual compuational details for evaluating the multitude of competing splits are presented later, with respect to a specific two-sample rank statistic (Wilcoxon). But, the attack and updating strategy apply to any two-sample *linear* rank statistic (Randles and Wolfe §9.1). Suffice it to say, for now, that the updating available makes for an $O(n)$ algorithm, that is as simple as the $O(n)$ LS algorithm. The story is not quite so simple when it comes to dealing with t statistic splitting and even less so for the censored analogues of the TSRS. Nevertheless, efficient algorithms can be devised

with the right organization. These are outlined below.

4. Extension to censored response. The original motivation for changing the splitting criterion was to enable tree techniques to be used for censored data. In fact, the notion of using a censored analogue of TSRS in a tree context had been advocated by Ciampi and Hogg (1982), but only in the situation where the predictors were dichotomous and no pruning algorithms were proposed. The carry-over to censored data is straightforward: the only change is the replacement of the TSRS with a censored version. Any member of the Tarone-Ware class (which includes both the Gehan and Mantel-Haenzel) was allowed as a user-specified option. Miller (1981) or Tarone and Ware (1977) have details on the statistics and their properties.

The computational ease alluded to in 3 above is now detailed with reference to splitting based on the two-sample Wilcoxon statistic. Again, suppose a split s partitions a node t into t_L and t_R. We now take $\phi(s,t)$ to be the value of the Wilcoxon statistic where t_L constitutes the first sample and t_R the second. The split s^* that maximizes this ϕ can be viewed as best separating the samples. The following pseudo-code demonstrates the ease of the update and the manner in which the best possible split for a given node is found. The first loop simply determines the mean and variance of the Wilcoxon for each possible split.

Wilcoxon Splitting Algorithm:

For each node t

Initialize BestStat = BestPred = BestSplitPoint = 0

Loop over all possible $N(t_L)$ values (left sample sizes)

$m \leftarrow N(t_L)$

$n \leftarrow N(t_R) = N(t) - N(t_L)$

$\text{Ave}(m) \leftarrow \frac{m(m+n+1)}{2}$

$\text{Var}(m) \leftarrow \frac{mn(m+n+1)}{12}$

End

Loop over all p predictors: $X_1, X_2, \cdots, X_p$

Initialize the rank sum: $R\text{Sum} \leftarrow 0$

Sort the node with respect to $X_{current}$

Reorder the Y's according the sorted $X_{current}$ values

Attach ranks to the Y's: $R(m) \leftarrow \text{rank}(Y_m)$

Loop over all potential split points which is equivalent to incrementing m

$R\text{Sum} \leftarrow R\text{sum} + R(m)$

TwoSam $\leftarrow (R\text{Sum} - \text{Ave}(m))/\sqrt{\text{Var}(m)}$

If abs(TwoSam) > BestStat **Then**

BestStat $\leftarrow$ TwoSam

BestPred $\leftarrow X_{current}$

BestSplitPoint $\leftarrow m$

etc.

End If

End

End

Important but simple issues such as not splitting on tied predictor values and efficient means for sorting and ranking have not been highlighted for clarity.

The same strategy would be used for splitting based on two-sample linear rank statistics. The form of such a statistic is

$$S = \sum_{j=1}^{m} a(R(j))$$

Here $R(\cdot)$ is as above - the rank assigned to $Y(\cdot) \in t_L$ where the ranking itself is done with respect to $Y(\cdot) \in t = t_L \cup t_R$. The $a(\cdot)$s are scores satisfying a nondecreasing and nonconstant condition, namely,

$$a(1) \leq \ldots \leq a(N(t)), \quad a(1) \neq a(N(t))$$

The Wilcoxon corresponds to using $a(j) = j$. The null expectation and variance for S are

$$E_0[S] = m\bar{a} \quad \text{Var}_0[S] = \frac{mn}{(m+n)(m+n-1)} \sum_{j=1}^{m+n} (a(j) - \bar{a})^2$$

where $\bar{a}$ is the average of the scores. Thus the only modifications to the above algorithm are (i) multiplying $Ave(m)$ and $\mathrm{Var}(m)$ by constants in the first loop and (ii) using $RSum \leftarrow sum + a(R(m))$ in the inner loop. The choice of the scores $a(\cdot)$ is left as a user specified option. If we make distributional assumptions, then Randles and Wolfe present formulae for *optimal* scores and expected scores, in the context of locally most powerful rank tests.

An issue that warrants comment here is the usage of large sample approximations in comparing the competing splits. That is, we have *standardized* TwoSam above, as opposed to ordering the statistics in terms of exact significance achieved (the best split corresponding to the greatest significance / smallest P value). There are compelling reasons for proceeding in this manner:

1. Exact significance results will only be available for small samples. Hollander and Wolfe (1973) tabulate P values only for $m + n < 25$.
2. Even in the instances where exact values are available it is difficult and inefficient to *automate* comparisons amongst competing splits.
3. We are not in the least interested in actual P values. All that is of consequence is the ordering of the statistics so that the best split can be ascertained.
4. Item 3 notwithstanding, the convergence to normality of such rank tests is very rapid. Such tests are intimately related to permutation tests (Randles and Wolfe §11.1) and the latter are known to be approximately normal for sample sizes as small as 10. Thus the ordering in 3 is being performed on appropriate quantities.

The implementation for the two-sample t is equally simple. The statistic itself is

$$T = (\mu_L - \mu_R)/\sqrt{\frac{s_L^2}{m} + \frac{s_R^2}{n}}$$

where μ_L and μ_R are means for the left and right subsamples and likewise s_L^2 and s_R^2 are the left respectively right unbiased estimates of variance. Also $m = N(t_L)$ and $n = N(t_R)$ as above. So we again take $\phi(s,t)$ to be the value of the statistic. As before s splits t into t_L and t_R and we seek the split that maximizes ϕ. There is no immediate way to update T itself corresponding to updating the split point (incrementing m and decrementing n). However, it is straightforward to update the mean and variance for the left subsample and correspondingly downdate the

right subsample quantities and then recompute the new T. The variance up and downdating was achieved using formula from Chan, Golub and LeVeque (1983). If, analogously to the weighted rank statistics presented above, a weighted t statistic was desired then the formulae in West (1979) for up and downdating weighted sums of squares could be used.

The form of the two-sample statistics for censored response (Gehan, Mantel-Haenszel, Tarone-Ware) are found in Miller Ch. 4. Details on the algorithm implemented for constructing a regression-tree using such statistics are available from the author. As with t statistic splitting, there is no immediate update for the statistics. Here efficiency is achieved by effectively managing the updating of the constituent quantities of the statistics (risksets, sample membership indicators). Gordon and Olshen (1985) also pursue tree-structured schemes when censoring is present. Their splitting criteria involve measures of distance between Kaplan-Meier survival curve estimates and certain point masses. No analyses are presented and so comparisons between the methods are precluded.

4. Revised Pruning Strategies.

An important difference between LS (or LAD) splitting as outlined in section 2 and any two-sample statistic splitting (section 3) is that the former provides a within node estimate of error viz. $SS(t)$: the within node sum of squares. Such is not the case for two-sample statistic splits, which only afford a measure of goodness of split. In general, these measures *cannot* be decomposed to attribute a within node error. This is consequential, since the within node errors form a key component of the pruning algorithm advocated in CART Ch. 3. The algorithm, therefore, does not carry over to the present situation and inasmuch as tree size is a fundamental issue, alternate approaches must be sought.

Initially, the focus of the modifications to the regression-treemethodology, was to facilitate a tree schema suitable for analysing censored data. Thus, the notion of using two-sample statistics as splitting rules was by no means sacred. In order to circumvent the problems posed by the absence of within node loss, an attempt to revert back to the original criteria was made. Specifically, LAD splitting was tried. Using LS was not entertained because of the unstable nature of the mean when estimated from survival curves. The intention was to account for the censoring by using medians based on the Kaplan-Meier survival curve estimated for each

node and then consider absolute deviations about these. There were a variety of difficulties associated with this attack that rendered it useless:

(a) Computationally this method was very slow. Even for LAD splitting in the uncensored response context, the program can be very slow unless some update algorithm for evaluating the absolute deviations is available (see §8.11.3). Whilst such an algorithm was written for the uncensored case, it appears too cumbersome for deviations about Kaplan-Meier medians.

(b) The actual splits obtained using this criterion on simulated data with known structure were not convincing. The method did not uncover the important variables or split points.

(c) The hope behind resurrecting LAD splitting was the inheritance of the pruning algorithm used in standard CART. However, even this did not materialize. An unstated necessity for the minimal cost-complexity algorithm (§3.3) to work is that the splitting criteria be convex. Let $\nu(t)$ be any sample median for the node t. Then in the uncensored case we have

$$\sum_{x_n \in t} |y_n - \nu(t)| \geq \sum_{x_n \in t_L} |y_n - \nu(t_L)| + \sum_{x_n \in t_R} |y_n - \nu(t_R)|$$

But this does *not* hold for censored ys and $\nu(\cdot) \stackrel{\text{def}}{=}$ the Kaplan-Meier median.

In fact, using deviations or functions thereof, will always be problematic when one of the quantities being differenced is subject to censoring.

The next attempt also involved trying to inherit the standard pruning algorithm. This time the tactic was to pursue a within node error measure whilst persevering with rank based splitting. The device was to use *one*-sample rank statistics (for example, Wilcoxon Signed Rank) as a method for generating measures for goodness-of-split. But before an attempt was made with respect to censored data it was deemed necessary to determine the performance of such a criterion on uncensored data. Again, the method failed. Each of the criticisms levelled in (a), (b) and (c) above still applied. The poor performance of using one-sample statistics in this manner can be understood simply in terms of what the statistic tests, namely, symmetry. It is possible to have within node symmetry but not homogeneity. So there is no a priori reason to expect this sort of splitting to do well. However, a worthwhile spinoff was the development

of a one-sample analogue to the Wilcoxon Signed Rank for censored data; see Segal (1985).

So what was needed was an altogether different tack. It was decided to preserve the concept of initially growing a very large tree and subsequently pruning this. What was sacrificed was the selection of a particular tree from the generated sequence by cross-validation. Further, the minimal cost-complexity pruning algorithm itself was replaced with some new pruning schemas.

The loss of cross-validation as a selection mechanism was not tragic. While the method had performed well its usage had several recognized flaws. The more detracting of these include: (i) inaccuracies and instabilities of the cross-validation estimates §8.7 and (ii) failure of the tree selected as optimal to preclude *noisy* splits §8.6. Indeed, the authors of CART promote user selection of the right-sized tree §3.4.3, §6.2. This should be done in an exploratory fashion and aided by the incorporation of subject matter knowledge.

But for such user selection, the user must be provided with a tree sequence and hopefully one that contains good candidate trees. It was to this end that the new pruning algorithms were created. Before expounding on these, it is important to reiterate what is being acquired from the CART approach – protection against the deficiencies of stopping rules, as highlighted in section 2. This protection derives solely from the tactic of growing a big tree and pruning it and is independent of the algorithm used to achieve this collapsing.

The first new method of pruning worked as follows. Recall that for each split we have stored the value of the statistic that led to that split.

- prescribe some threshold value (for comparison against the statistics)
- starting at the bottom of the tree, step up and
  - collapse nodes (make terminal) that arose from splits whose statistics did not surmount the threshold
  - retain nodes and all their ancestors where the improvement was sufficient i.e. the threshold was exceeded

Whilst the performance of this procedure in practise was passable, a couple of objections exist. Firstly, the prechosen cutoff - critical in determining the tree sequence - suffers from

all the rigidity associated with specifying such levels. Secondly, there is a tendency for very large branches to be retained. If, perhaps fortuitously due to small sample size, a split near the bottom of the tree achieved large significance then the entire branch leading to that split will be kept in *all* subsequent trees. The next method overcomes both these drawbacks.

The second technique implemented was motivated by a need to redress the above liabilities. Again the starting point is a very large tree. The basic idea was the following:

- sequentially examine every internal node, starting with the one associated with the least significant split statistic and proceeding through to the most significant
- examine the subtree (possibly null) emanating from the node under consideration i.e. look at all the descendent nodes from the current position in the tree
  - if the subtree contains no consequential splits, prune it away
  - alternatively, if there are worthwhile splits, preserve the subtree

This was reworked into the framework below:

- step up the tree, assigning to each internal node the *maximum* split statistic contained in the subtree of which the node under consideration is the root
- collect all these maxima and place them in increasing order
- the first pruned tree of the sequence corresponds to locating the highest node in the tree possessing the smallest maximum and removing all its descendents
- the second tree of the sequence is then obtained by reapplying this process to the first tree and so on until all that remains is the root node

This procedure is illustrated as part of the examples in the next section. The associated output is also displayed. Essentially, each internal node is linked with the maximum split statistic contained in the subtree for which the node is the root. The pruning sequence is then determined by the order of these maxima. This is equivalent to the look-ahead proposed in the above formulation but obviates the need to contrive a definition (in terms of thresholds) for *consequence*. In practise it has been found to perform well, one basis for assessment being that comparable tree sequences to those yielded by the CART technique of minimal cost-complexity pruning are extracted.

5. Examples.

The first example analysed by the new regression-tree techniques is the simulation model discussed in §8.6. By generating data in accordance with a known model it is possible to assess whether the methods are performing reasonably. It was important to establish this in a familiar setting, before leaping into the unknown world of censored response. We take $p = 10$ and impose that $X_1, X_2, \ldots, X_{10}$ be independent. Also

$$\Pr(X_1 = -1) = \frac{1}{2} = \Pr(X_1 = 1)$$

$$\Pr(X_j = -1) = \Pr(X_j = 0) = \Pr(X_j = 1) = \frac{1}{3},$$
$$j = 2, 3, \ldots, 10.$$

Let Z be independent of $X_1, X_2, \ldots, X_{10}$ and $Z \sim N(0, 2)$. Then if $X_1 = 1$ set

$$Y = 3 + 3X_2 + 2X_3 + X_4 + Z,$$

and if $X_1 = -1$ set

$$Y = -3 + 3X_5 + 2X_6 + X_7 + Z.$$

Variables X_8, X_9, X_{10} are noise. The learning sample comprized 200 cases generated from this model.

The example consists of two distinct regression equations, with the choice of equation dictated by the binary variable X_1. This should be reflected by having X_1 chosen as the first splitting variable. Then predictors should be emerge as splitting variables in order of the magnitude of their coefficients i.e. in the subtree corresponding to the top equation we would expect X_2 to enter ahead of X_3 which in turn would enter before X_4 and similarly for the subtree corresponding to the bottom equation. Examination of the tree diagram (Figure 1) reveals that everything is as it should be. This tree and the associated sequence (Table 1) were obtained using Wilcoxon splitting. However, analogous results are achieved using t statistic splitting and these in turn agree with the output obtained from LS or LAD. The way to read Figure 1 is as follows: the cases for which the condition below a given node holds true go down the left branch of the diagram and those for which the condition is false go down the right. Thus, of the 42 cases contained in node 4, the 29 having X_6 values of -1 or 0 get assigned to

node 8 and the remaining 13, whose X_6 value is 1 go to node 9. The squares represent terminal nodes and the numbers below these are the medians of the Ys in that node.

The first split is on the binary variable X_1 and separates the 94 cases generated from the bottom equation from the 106 cases generated from the top equation. Then the left side splits repeatedly on X_8 and X_6 with one split on X_7. The right subtree splits repeatedly on X_2 and X_3 with one split on X_4. All of these splits occur in the correct order. Note how user selection of the "optimal" tree enables the devious exclusion of noisy splits. The tree so chosen - number 18 in Table 1 - has 14 terminal nodes, which agrees with the range obtained by conventional CART of 12 to 16. Also note the tendency for the pruning process to take off two terminal nodes at a time (indicated by the total number of terminal nodes decreasing by one). This again is in accordance with CART regression findings and is explained in §8.5. Note that the node numbers quoted as subtree roots pertain to the initial large tree, so that some of them cannot be interpreted from the tree actually selected. Finally, the rationale for the apparent discrepancy between the size of the largest CART tree (200 terminal nodes) and the largest tree resulting from the two-sample method (32 terminal nodes), lies in the fact that differing minimum node size parameters were used.

Other uncensored examples were subject to analysis by both the new and old tree schemas. Illuminating aspects emerged from both analyses and the results reinforced the reasonableness of the new approach. The examples attacked were the Swiss Fertility (Mosteller and Tukey, 1977) and Boston Housing (Belsley, Kuh and Welsch, 1980) datasets; for details see Segal (1985).

The canonical example for illustrating the performance of any regression technique where the response is subject to censoring is the Stanford Heart Transplant data; see Miller (1976), Buckley and James (1979) and the more recent "nonparametric" treatments of Tibshirani (1984), Doksum and Yandell (1982) and Owen (1985) for instance. The celebrated Proportional Hazards Model, Cox (1972), has also been applied.

A brief data description is now given. The response Y is $\log_{10}$ survival time, where the survival time is the time (in days) until death due to rejection of the transplant heart. There are $p = 2$ predictors: X_1 the age of the recipient and X_2 a tissue mismatch score measuring

recipient and donor tissue compatability. 157 cases were analysed, there being a 35% censoring rate.

What has consistently emerged from the plethora of analyses is that age is the more significant predictor. Further, the nonparametric approaches have revealed a cutoff value of roughly 50 years, in that the subpopulations so defined (≤ 50 and ≥ 50 say) have distinct survival characteristics. A scatterplot of $\log_{10}$ survival against age is shown in Figure 2. The cutoff is not overly apparent to the naked eye, presumably because of the nonuniform censoring pattern.

Regression trees, using two-sample statistic splitting, were used to analyse the data. In particular, the Gehan statistic was used in conjunction with subtree maximal statistic pruning to produce both the tree schematic in Figure 3 and the tree sequence in Table 2. Figure 3 is to be read in the same manner as Figure 1, except now the values below the square terminal nodes are Kaplan-Meier medians. It is worth recording that neither the initial large tree, nor the pruned sequnce were substantially altered by using other splitting statistics from the Tarone-Ware class. In fact, the key first split was identical in the cases examined. What is immediately evident from the tree diagram is the confirmation of the previous findings. Firstly, age clearly emerges as the more consequential predictor (though see CART §5.3.4 for an automated means for predictor ranking that overcomes possible *masking* - this is not an issue here since there are only two predictors). Secondly, the cutoff at around 50 is reflected by the value of the first split point. The value of the statistic for this split is 4.91 and this, being a standardized z indicates the significance of the division.

However, the analysis can proceed further. A natural first summary for a terminal node when we have a censored response is the estimated Kaplan-Meier survival curve $\hat{S}$ (Kaplan and Meier, 1958). The program also provides the user with the possibility of extracting certain derived quantities, such as the Kaplan-Meier median $\hat{S}^{-1}(0.5)$ as node summaries or predictions. Figure 4 features superposed survival curves for each of the 4 terminal nodes. The curve corresponding to node 3 in the tree schematic of Figure 3 lies appreciably below the curves for nodes 4 and 6. Node 3 contains the ≥ 50 agegroup. Their survival prospects are noticeably worse than the bulk of the ≤ 50 group, contained in nodes 4 and 6, as would be expected. But,

the survival characteristics for node 7 resemble those for node 3. Node 7 contains patients who are middle-aged (41 - 50) as opposed to young and who also have high tissue mismatch scores. Thus it is not surprising that they possess equally poor survival prospects. It is the extrication of precisely such local interactions that make tree techniques so powerful. Of course caution must be exercised in interpreting survival curves based only on 8 cases. The tree structured approach affords many other compelling advantages. Some, such as the easily understood and interpretable nature of the output have hopefully emerged unstated from these examples. Others, such as the exploitation of information contained in missing values and purposeful dimensionality reduction, have not been illustrated. A partial itemization of tree virtues is compiled in CART §2.7.

6. Discussion.

Much remains to be done in refining and toward better understanding the regression-treetechniques introduced in this paper. One route being pursued, with a view to greater appreciation of the methodologies performance, is an extensive simulation study. This is particularly geared to the censored response situation. It is hoped to elucidate the effect of varying censoring distributions, establish what sorts of splits are selected and how "significant" these are. Additional real-world data sets are also being analysed. The performance of the trees to date has been very encouraging. Another line of attack involves determining the asymptotic behaviour of such models. CART §12.3 demonstrates the consistency of regression-trees under mild regularity conditions and it is projected that similar results hold for the new tree schemas. Indeed CART concludes by asserting that no theoretical justification has been obtained for any of the specific splitting rules used, pruning strategies or cross-validation. To the extent that these are the aspects that have been altered, consistency follows immediately. This even extends to the situation where the response is subject to censoring as asserted by Gordon and Olshen (1985).

One other extension being developed is to multi-response situations. The only envisaged change is the replacement of LS, say, with some multivariate analogue. The splitting criterion ϕ would then take the form

$$\phi(s,t) = SS(t) - SS(t_L) - SS(t_R)$$

where now

$$SS(\cdot) = \sum (\vec{y} - \mu(\cdot))' \, \mathbf{A} \, (\vec{y} - \mu(\cdot)).$$

The sum is over cases contained in the appropriate node and $\mu(\cdot)$ designates the mean for that node. A is a diagonal matrix used to weight predictors. Alternatively, the two-sample statistic splitting approach could be extended by, for instance, maximizing the Mahalanobis distance between the two subsamples. Nothing is sacrosanct about the covariance matrix, so other metrics could be used. Note, however, that due to the absence of an order in $\mathbf{R}^q$, for $q > 1$: the dimension of the response, rank methods are not applicable.

The thrust of CART was to develop a method that, despite not possessing any theoretical optimality properties or rich large-sample results, performed well in practise. The examples presented throughout the monograph establish that legitimacy. Likewise, the regression-treeprocedures involving rank based splitting rules have displayed their validity "in the field" and have the potential to constitute a good analysis tool.

7. References.

BELSLEY, D.A., KUH, E. and WELSCH, R.E. (1980) Regression Diagnostics, Wiley

BLOCH, D.A. and SEGAL, M.R. (1985) *Empirical Comparison of Approaches to Forming Strata: Using Classification Trees to Adjust for Confounding Variables*, submitted to JASA

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J. (1984) Classification and Regression Trees, Wadsworth

BUCKLEY, J. and JAMES, I.R. (1979) *Linear Regression with Censored Data*, Biometrika **66** 429 - 436

CART^{TM} (1984) Copyright, California Statistical Software, Inc.

CHAN, T.F., GOLUB, G.H. and LeVEQUE, R.J. (1983) *Algorithms for Computing the Sample Variance: Analysis and Recommendations*, The American Statistician **37** No. 3, 242 - 247

CIAMPI, A. and HOGG, S. (1982) *A Recursive Partition Algorithm for the Classification of Survival Data*, Unpublished manuscript

COX, D.R. (1972) *Regression Models and Life Tables*, JRSS B **34**, 187 - 202

DOKSUM, K A. and YANDELL, B.S. (1982) *Properties of Regression Estimates Based on Censored Survival Data*, Festschrift for Erich L. Lehmann, Edited by P.J. Bickel, K.A. Doksum and J.L. Hodges, Jr.

FRIEDMAN, J.H. (1979) *A Tree Structured Approach to Nonparametric Multiple Regression*, In Smoothing Techniques for Curve Estimation, Edited by T. Gasser and M. Rosenblatt. Springer-Verlag.

GORDON, L. and OLSHEN, R.A. (1985) *Tree Structured Survival Analysis*, Unpublished manuscript

HOLLANDER, M. and WOLFE, D.A. (1973) Nonparametric Statistical Methods, Wiley

KAPLAN, E.L. and MEIER, P. (1958) *Nonparametric Estimation from Incomplete Observations*, JASA **53**, 457 - 481

MILLER, R.G. Jr. (1981) Survival Analysis, Wiley

MORGAN, J.N. and SONQUIST, J.A. (1963) *Problems in the Analysis of Survey Data, and a Proposal*, JASA **58**, 415 - 434

MOSTELLER, F. and TUKEY, J.W. (1977) Data Analysis and Regression, Addison Wesley.

OWEN, A.B. (1985) *Ph.D Thesis in preparation*, Department of Statistics, Stanford University

RANDLES, R.H. and WOLFE, D.A. (1979) Introduction to the Theory of Nonparametric Statistics, Wiley

SEGAL, M.R. (1985) *Ph.D Thesis in preparation*, Department of Statistics, Stanford University

TARONE, R.E. and WARE, J. (1977) *On Distribution-Free Tests for Equality of Survival Distributions*, Biometrika **64**, 156 - 160

TIBSHIRANI, R.J. (1984) *Local Likelihood Estimation*, Ph.D Thesis, Department of Statistics, Stanford University

WEST, D.H.D. (1979) *Updating Mean and Variance Estimates: An Improved Method* Communications of the ACM, **22**, no. 9 532-535

Tree Diagram for Simulation Model

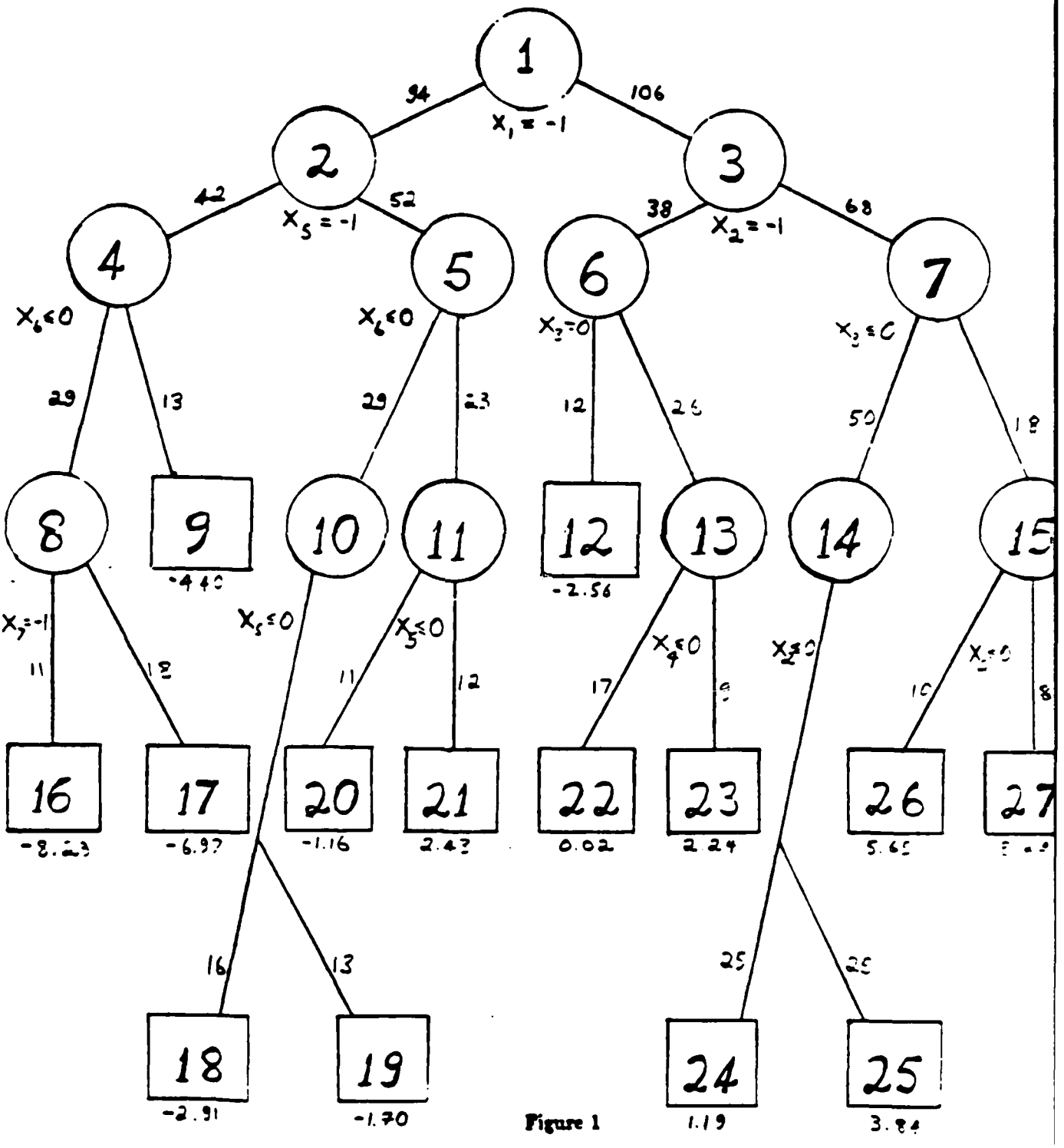


Figure 1

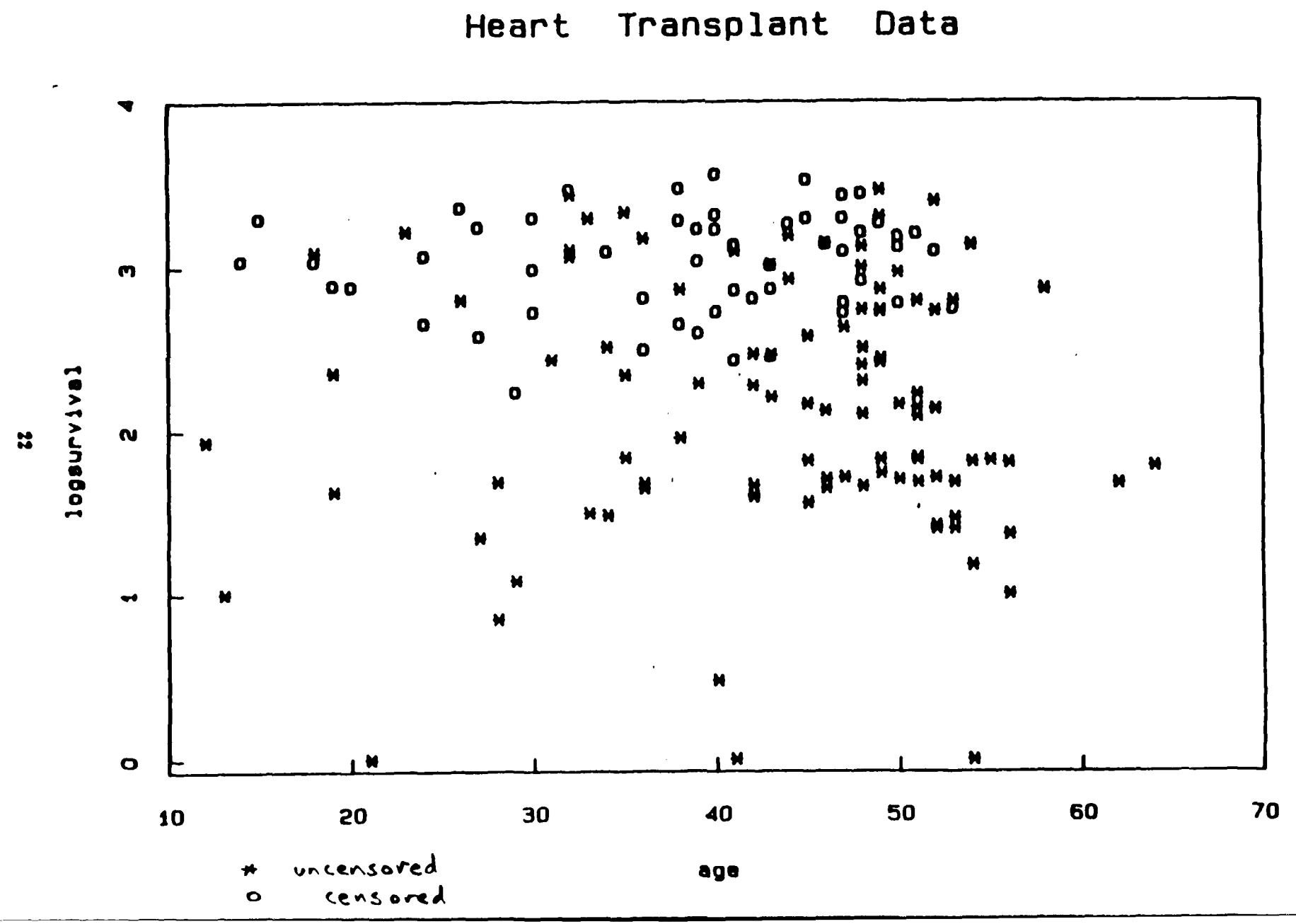
Table 1

Pruning by Subtree Maximal Statistics

Simulation Study

Tree Number	Terminal Nodes	Subtree Root Node	Subtree Maximal Statistic
1	32	-1	0.00
2	31	30	0.73
3	30	47	1.04
4	29	34	1.14
5	28	36	1.27
6	27	60	1.38
7	26	21	1.40
8	25	9	1.42
9	24	12	1.43
10	23	22	1.44
11	22	46	1.76
12	21	26	1.78
13	20	17	1.86
14	18	29	2.01
15	17	20	2.05
16	16	28	2.26
17	15	16	2.29
18	14	15	2.40
19	13	14	2.49
20	12	13	2.61
21	11	8	2.65
22	10	10	2.66
23	8	11	3.21
24	7	4	3.52
25	6	6	3.68
26	5	5	4.00
27	4	7	4.50
28	3	3	6.20
29	2	2	6.59
30	1	1	9.29

Figure 2


Heart Transplant Data
logsurvival
0
1
2
3
4
10
20
30
40
50
60
70
age
* uncensored
o censored

Tree Diagram for Stanford Heart Transplant Data

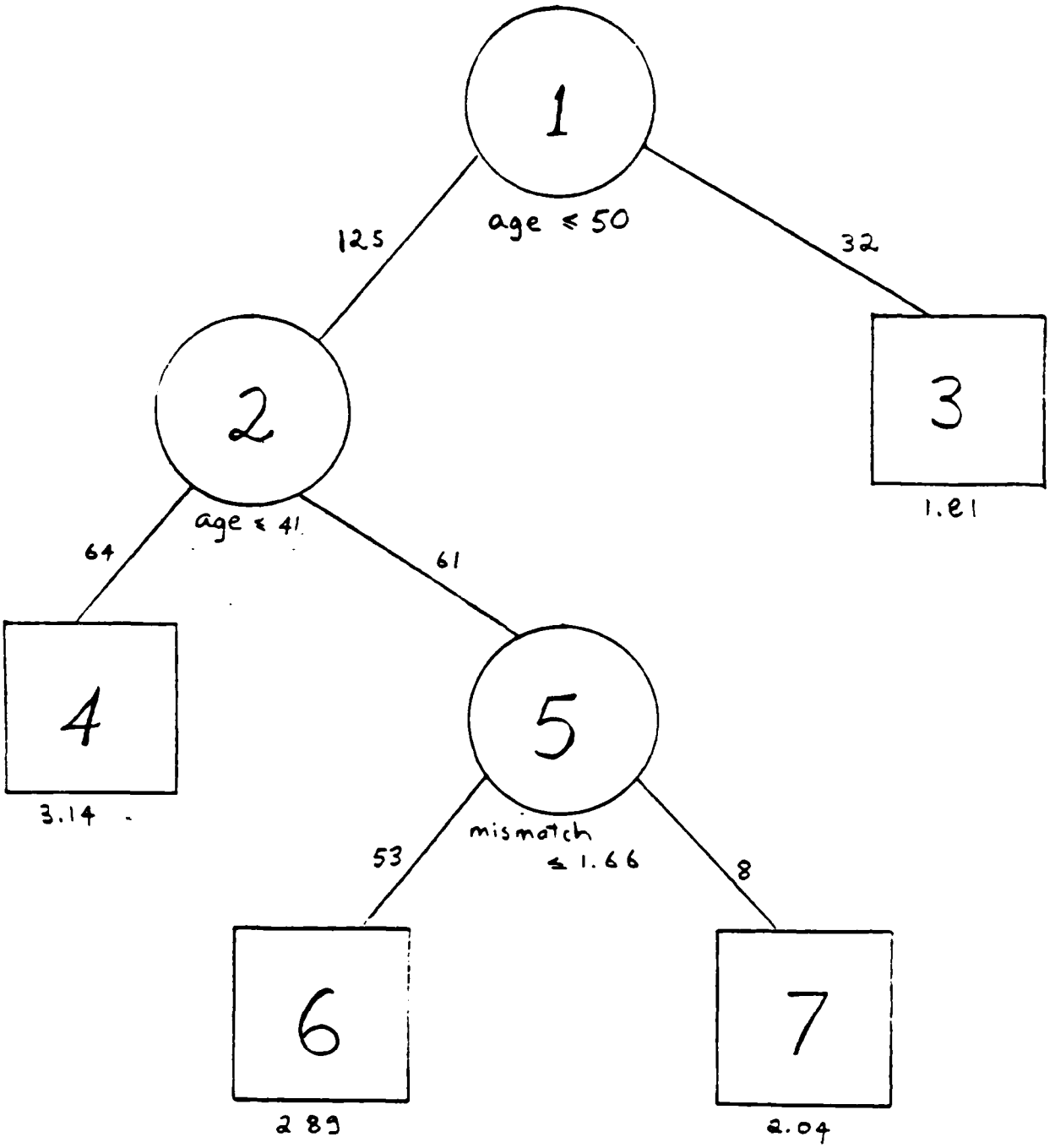


Figure 3

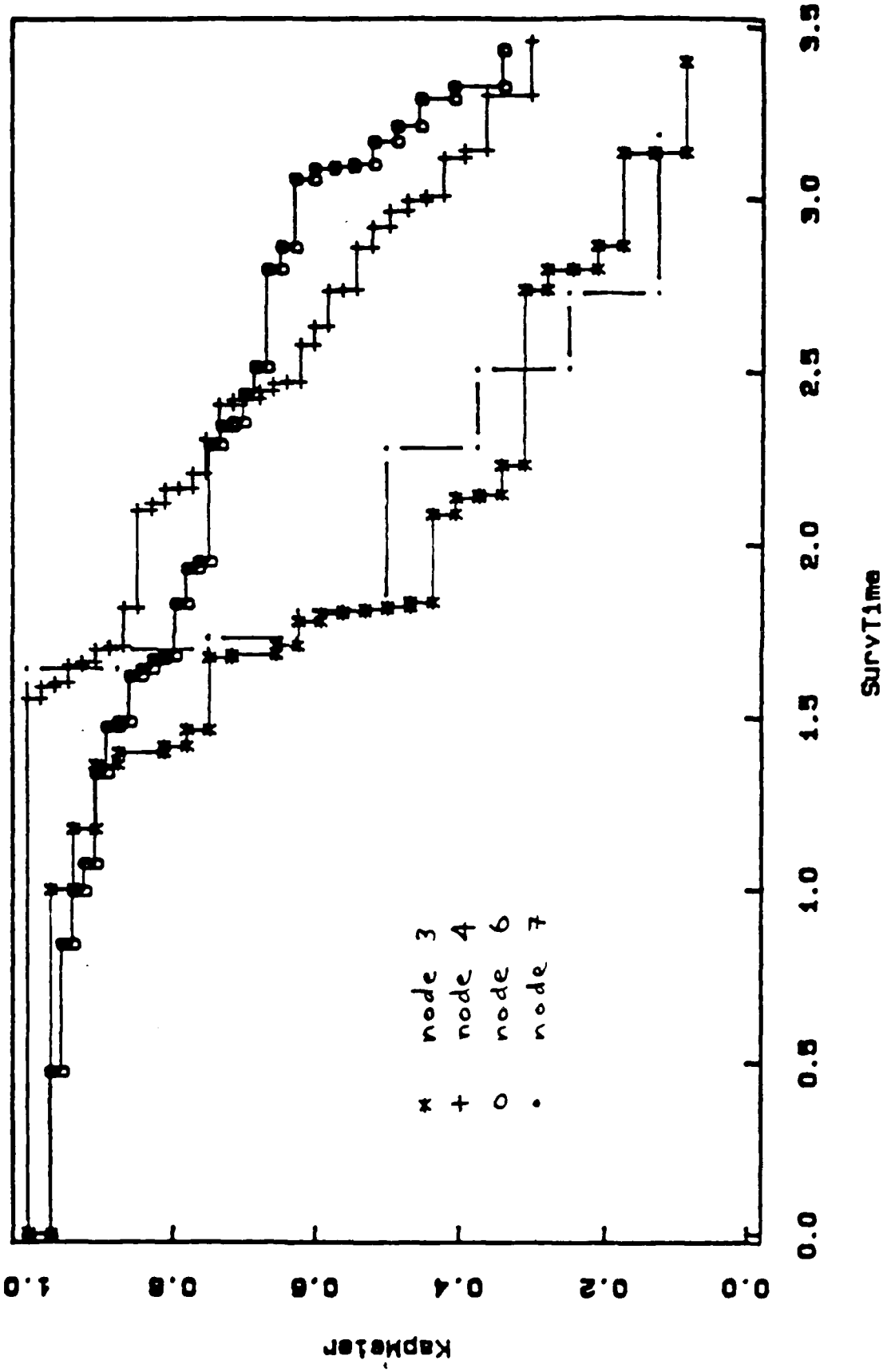

Kaplan Meier Curves for Terminal Nodes
KapMeier
SurvTime
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
2.0
2.5
3.0
3.5
* node 3
+ node 4
o node 6
. node 7

Figure 4

Table 2

Pruning by Subtree Maximal Statistics

Stanford Heart Transplant Data

Tree Number	Terminal Nodes	Subtree Root Node	Subtree Maximal Statistic
1	14	-1	0.00
2	13	14	0.60
3	12	10	0.99
4	11	6	1.17
5	10	16	1.25
6	9	3	1.79
7	6	13	2.43
8	4	8	2.48
9	2	2	2.79
10	1	1	4.91

1. REPORT NUMBER	2. GOVT ACCESSION NO	3. RECIPIENT'S CATALOG NUMBER
LCS 15	AD-A158546	

4. TITLE (and Subtitle)	5. TYPE OF REPORT & PERIOD COVERED
RECURSIVE PARTITIONING USING RANKS	TECHNICAL
	6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)	8. CONTRACT OR GRANT NUMBER(s)
Mark Robert Segal	N00014-83-K-0472

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Department of Statistics and Computational Group Stanford Linear Accelerator Center Stanford University, Stanford, CA 94305	

11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE
U.S. Office of Naval Research Department of the Navy Arlington, VA 22217	August 1985
	13. NUMBER OF PAGES 25

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report)
	UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Navy position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Regression-Tree, Splitting Rule, Pruning, Censoring,
Two-Sample Rank Statistic

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Replacing the conventional splitting rules used in constructing regression trees by rules based on two sample rank statistics affords many advantages and equally poses some problems. Among the former are (1) computational ease, (2) invariance under monotone transformations of the response and (3) worthwhile extension to censored data. The difficulties involve devising good pruning strategies in the absence of within node loss. These are addressed using look-ahead, bottom-up techniques. Some real-world and simulation performances of the methodology are presented.

END

FILMED

10-85

DTIC